REGRESJA DO ŚREDNIEJ JAKO PODSTAWOWE ZAGROŻENIE DLA BADANIA ZMIANY WYNIKÓW SKRAJNYCH

Nie ma chyba techniki statystycznej, której terminologia i rozumienie przechodziłyby tak osobliwe koleje losu, jak to miało miejsce w wypadku analizy regresji. Technika ta nie ma nic wspólnego z tym, co sugeruje jej nazwa (łac. regressio – cofanie się, ruch wstecz – regres, redukcja, dezorganizacja, cofanie się; Tokarski, 1980). Termin „regresja" został, jak wiadomo, przez kolejne pokolenia statystyków i badaczy przejęty od Galtona, na określenie techniki statystycznej służącej do predykcji zmiennych na podstawie znajomości innych zmiennych, chociaż Galton pierwotnie rozumiał przez nią zupełnie co innego – domniemaną siłę biologiczną sprawiającą, że rozmiary osobników potomnych są bliższe średniej niż rozmiary rodziców. Nazewnictwo metod i jej proweniencja jest jednak sprawą drugorzędną, ważniejszy jest fakt, że analiza regresji, jedna z najważniejszych metod w naukach biologicznych, ekonomicznych i społecznych, jest badaczom dobrze znana. W przeciwieństwie natomiast do tego, zjawisko, którego odkrycie dało początek skonstruowaniu metod badania związków między zmiennymi, w tym analizy regresji, określane dziś jako regresja do średniej (*regression toward the mean*), pozostaje zbyt mało znane, zwłaszcza jeśli wziąć pod uwagę rozmiary szkód powodowanych przez jego nieuwzględnianie.

Regresja do średniej polega, najogólniej mówiąc, na tym, że skrajne grupy obserwacji mają przy drugim testowaniu wyniki bliższe średniej niż przy pierwszym testowaniu. Innymi słowy, badani, którzy za pierwszym razem osiągnęli wyniki poniżej średniej, będą mieli przy drugim pomiarze wyniki bliższe średniej (czyli wyższe niż za pierwszym razem), natomiast badani, którzy przy pierwszym pomiarze mieli wyniki powyżej

średniej, przy drugim pomiarze osiągną wynik również bliższy średniej (czyli niższy). Stanie się tak nawet wtedy, kiedy manipulacja oddzielająca oba pomiary była całkowicie nieskuteczna, albo kiedy w ogóle nie było żadnej manipulacji (Nesselroade, Stigler i Baltes, 1980; Campbell i Kenny, 1999).

Co to oznacza dla badań empirycznych, na przykład badań dotyczących zmiany osobowości albo skuteczności psychoterapii, nie trzeba podkreślać. Jeśli na przykład badacz zaprojektował metodę terapii zaburzeń lękowych i do badań sprawdzających jej skuteczność wybrał na podstawie testu lęku osoby bardzo lękowe, to po zakończeniu terapii i powtórnym zastosowaniu testu lęku badacz ten odkryje, że w badanej grupie lęk jest teraz niższy niż w pierwszym pomiarze. Co więcej, najsilniejszą „poprawę" wykażą osoby, które w pierwszym pomiarze były najbardziej lękowe. Stanie się tak nawet wtedy, jeśli terapia była zupełnie nieskuteczna.

Niniejszy artykuł jest poświęcony temu właśnie zjawisku. Zawarto w nim genezę terminu „regresja do średniej", wyjaśnienie możliwych jej przyczyn, przykład badań, których wyniki mogły być spowodowane regresją do średniej oraz opis niektórych sposobów, poprzez które można się ustrzec wynikających z niej błędów lub przynajmniej złagodzić jej skutki.

❑ 1. Geneza terminu „regresja do średniej"

Terminu „regresja" po raz pierwszy użył Francis Galton (1886) w słynnym artykule zatytułowanym „Regression towards mediocrity in hereditary stature". W tekście tym Galton na wstępie przypomniał swoje eksperymenty dotyczące dziedziczenia rozmiarów ziaren groszku. Z badań tych wynikało, że rośliny rozwinięte z dużych ziaren wytwarzają ziarna mniejsze niż te, z których same pochodzą, natomiast rośliny rozwinięte z małych ziaren wytwarzają ziarna większe. Galton przedstawił następnie wyniki podobnych badań, ale dotyczących wzrostu u ludzi. Badania te zostały przeprowadzone na 205 parach rodziców i ich 928 dorosłych dzieciach. Wzrost każdej matki został przemnożony przez 1,08 (ponieważ mężczyźni są średnio wyżsi od kobiet o 8%), a następnie obliczono średnią wzrostu obojga rodziców. W ten sposób Galton uzyskał jeden wskaźnik wzrostu dla obojga rodziców i nazwał go średnim wzrostem rodziców (*mid-parentage height*). Porównując uśredniony wzrost rodziców ze wzrostem ich dzieci, Galton zauważył, że wysocy rodzice mają dzieci średnio niższe od siebie, natomiast niscy rodzice mają dzieci wyższe od siebie.

Analizując te wyniki, Galton nie zdawał sobie jeszcze sprawy, że ma do czynienia z artefaktem statystycznym i swoje wyniki interpretował w kategoriach biologicznych. Przypuszczał mianowicie, że na wzrost dzieci wpływa wzrost nie tylko ich biologicznych rodziców, lecz także genera-

cji poprzednich. Jeśli nawet rodzic był bardzo wysoki, a jego przodkowie mieli różny wzrost, to dziecko było niższe niż rodzice, ponieważ, zdaniem Galtona (1886), wzrost dzieci jest wypadkową wzrostu rodziców i generacji wcześniejszych. Zjawisko to nazwał Galton „dziedziczną regresją do przeciętności" (*filial regression towards mediocrity*).

Takie biologiczne wyjaśnienie wydawało się logiczne, przynajmniej na podstawie ówczesnej wiedzy o dziedziczeniu, było jednak zupełnie błędne. Galton zdał sobie z tego sprawę po kilku latach, kiedy zauważył, że rodzice wysokich dzieci są niżsi od tych dzieci, a rodzice niskich dzieci są od nich wyżsi (Stigler, 1997). Regresja do średniej nie mogła więc w tym przypadku mieć nic wspólnego z przyczynowością ani genetyczną, ani żadną inną, ponieważ rodzice nie zwykli dziedziczyć wzrostu po dzieciach, i żadne w ogóle przyczyny nie działają wstecz w czasie. Wzrost jest oczywiście cechą w dużym stopniu dziedziczną, więc wzrost rodziców jest oczywiście przyczyną wzrostu dzieci, ale czynniki genetyczne nie są przyczyną regresji wyników do średniej. Niemniej jednak, termin „regresja do średniej" upowszechnił się w statystyce, co więcej, terminem „regresja" zaczęto określać technikę statystyczną pozwalającą przewidywać wyniki jednej zmiennej na podstawie innych zmiennych.

❑ 2. Przyczyny zjawiska regresji do średniej

Regresja do średniej występuje, ponieważ dane empiryczne obciążone są błędem, polegającym na losowych fluktuacjach wyników pomiarów poszczególnych osób w drugim pomiarze w porównaniu z pierwszym. Jak wiadomo, na wynik otrzymany w teście składają się dwa elementy: wynik prawdziwy oraz składnik błędu. Wyniku prawdziwego badanej osoby nie znamy, znamy tylko wynik otrzymany. Nie znamy też dokładnie składnika błędu, lecz możemy założyć, że działa on w sposób losowy, to znaczy jednakowe jest prawdopodobieństwo zawyżenia bądź zaniżenia wyniku wskutek błędów pomiaru (jeśli prawdopodobieństwo to nie jest jednakowe, to nie mówimy już o składniku błędu, tylko o błędzie systematycznym narzędzia, który nie ma związku ze zjawiskiem regresji do średniej).

Fakt, że na wynik otrzymany składa się, oprócz wyniku prawdziwego, także błąd pomiaru, ma doniosłe konsekwencje, kiedy do badań wybierane są tylko osoby o wysokich lub niskich wynikach. Wybór ten bowiem dokonywany jest na podstawie nie tylko wyniku prawdziwego, lecz także błędu pomiaru. Jeśli na podstawie narzędzia obciążonego błędem wybieramy osoby o niskich wynikach, to uzyskujemy grupę, w której znajdują się również osoby, których obniżone wyniki są spowodowane tylko błędem pomiarowym narzędzia. Rozważmy następujący przykład: pewna osoba ma wynik otrzymany, który jest równy dwóm odchyleniom standardowym poniżej średniej. Istnieją trzy możliwości: (1) wynik prawdziwy tej

osoby jest równy dwóm odchyleniom standardowym poniżej średniej; (2) wynik prawdziwy jest jeszcze niższy niż dwa odchylenia standardowe, a błąd pomiaru zawyżył wynik i (3) wynik prawdziwy to mniej niż dwa odchylenia standardowe w dół, zaś błąd pomiaru obniżył wynik. Pomijając chwilowo ewentualność pierwszą, jako zdarzającą się najrzadziej, zastanówmy się, która z pozostałych możliwości – (2) lub (3) jest bardziej prawdopodobna (a tym samym zdarza się częściej). Częstsza będzie możliwość trzecia, a to dlatego, że z rozkładu normalnego wynika, że wynik jest tym rzadszy, im bardziej odległy od średniej. W konsekwencji, powtórny pomiar przyniesie wyniki średnio wyższe niż za pierwszym razem, ponieważ bardzo mało prawdopodobne jest, aby błąd ponownie obniżył wyniki wszystkim badanym, u których je obniżył w pierwszym pomiarze.

Podobnie stanie się z grupą osób o wysokich wynikach. Jeśli na przykład badacz zamierza przetestować nową metodę terapii depresji, i w tym celu rekrutuje do badań osoby osiągające w jakiejś skali depresji wynik powyżej przyjętego progu, to kryterium wyboru osób jest ich wynik otrzymany, a nie wynik prawdziwy. Znaczy to, że uzyskujemy grupę, w której znajdują się również osoby z podwyższonymi wynikami spowodowanymi tylko błędem pomiarowym narzędzia. Jest mało prawdopodobne, aby w wypadku tych samych osób w drugim pomiarze zdarzył się taki sam błąd – oczekujemy, że u połowy osób błąd będzie powodował, że wyniki skierują się w przeciwną stronę. Powtórny pomiar całej grupy przyniesie więc średnią bliższą ich wynikowi prawdziwemu, a tym samym niższą niż za pierwszym razem.

Zjawisko to można przeanalizować dokładniej, posługując się tabelą zaproponowaną przez Streinera (2001).

Tabela 1

Podział badanych ze względu na wynik prawdziwy i wynik otrzymany

Grupa	Wynik prawdziwy	Wynik otrzymany	Decyzja	Wynik retestu
A	Poniżej kryterium	Poniżej kryterium	Nie włączać	brak
B	Poniżej kryterium	Powyżej kryterium	Włączyć	Powyżej kryterium Poniżej kryterium
C	Powyżej kryterium	Powyżej kryterium	Włączyć	Powyżej kryterium Poniżej kryterium
D	Powyżej kryterium	Poniżej kryterium	Nie włączać	brak

Źródło: Streiner, 2001.

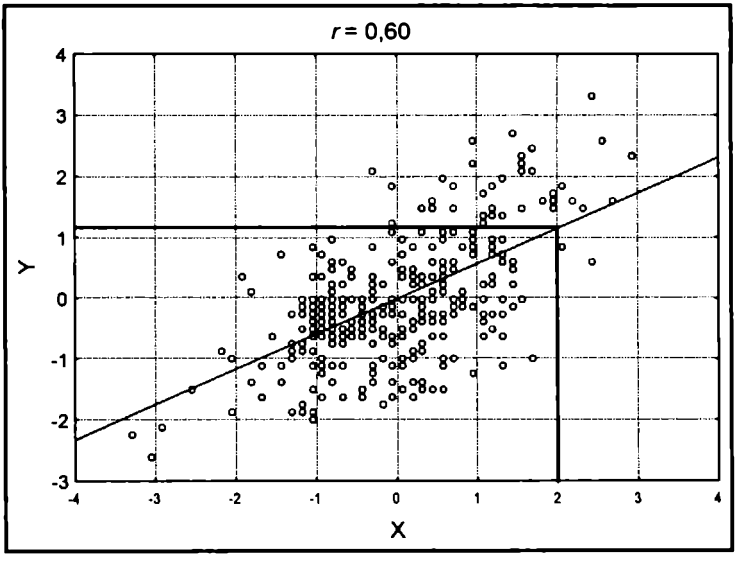
Tabela ta przedstawia cztery grupy osób badanych, wyróżnione ze względu na ich wynik prawdziwy oraz wynik otrzymany w pierwszym testowaniu. Badania dotyczyły terapii depresji; wynik powyżej kryterium wskazuje na depresję. Grupy A i B mają wynik prawdziwy poniżej kryterium (czyli nie zostałyby włączone do badań, gdyby badacz miał wgląd w wynik prawdziwy), grupy C i D mają wynik prawdziwy powyżej kryterium (zatem zostałyby włączone do badań, gdyby badacz miał dostęp do wyniku prawdziwego). Wskutek błędu pomiaru, spośród dwóch grup, które mają wynik prawdziwy poniżej kryterium, grupa A ma wynik poniżej kryterium (nie zostaje więc włączona do badań), natomiast grupa B ma wynik powyżej kryterium, więc zostaje do badań włączona. Podobnie, spośród grup, które mają wynik prawdziwy powyżej kryterium, grupa C zostanie – na podstawie wyniku otrzymanego – włączona do badań, grupa D natomiast do badań nie trafi (Streiner, 2001).

Załóżmy, że terapia depresji, która jest przedmiotem tych badań, jest całkowicie nieskuteczna. Wtedy podczas retestu wszystkie osoby z grupy C będą miały taki sam wynik prawdziwy, jak w pierwszym badaniu, natomiast część z nich będzie miała wynik otrzymany poniżej kryterium. Stanie się tak, ponieważ błąd pomiaru ma średnią zero i rozkład normalny, co oznacza, że w połowie pomiarów zawyża wynik, a w połowie go obniża. Zatem, w powtórnym badaniu, część osób będzie miała wyniki niższe i trafi do grupy „wyleczonych", nawet jeśli interwencja terapeutyczna była całkowicie nieskuteczna, czyli wyniki prawdziwe tych osób nie uległy zmianie (Streiner, 2001). Najprościej mówiąc, regresja do średniej wynika z faktu, że w grupie osób uzyskujących dobre wyniki (cokolwiek to znaczy) znajdą się osoby, które swój dobry wynik zawdzięczają szczęściu, a wśród osób o kiepskich wynikach będą takie, które miały po prostu pecha. Niewielkie są szanse, aby szczęście powtórzyło się u wszystkich za drugim razem (co wynika z samej definicji szczęśliwego trafu), podobnie niewielkie są szanse, aby za drugim razem mieć pecha. Podobne zjawisko zajdzie w grupie B. Badacz dojdzie do wniosku, że terapia depresji jest skuteczna, ponieważ duża część osób została „wyleczona" z depresji (Streiner, 2001).

W powyższym przykładzie mieliśmy do czynienia z sytuacją, w której analizy polegały na powtórnym pomiarze tej samej zmiennej. Inna z możliwych manifestacji zjawiska regresji do średniej wiąże się z analizami natury korelacyjnej. Ogólnie mówiąc, dla dowolnego równania regresji, które zostało wyznaczone na podstawie nieidealnej korelacji, wynik przewidywany dla zmiennej zależnej, dla danej wartości predyktora, jest bliższy średniej niż wartość predyktora. Innymi słowy, dla dowolnych dwóch nieidealnie skorelowanych zmiennych, osoby, które mają skrajny wynik w jednej z tych zmiennych, będą miały w drugiej z nich oczekiwany wynik nie tak bardzo skrajny (Campbell i Kenny, 1999). Jeśli bowiem zmienne skorelowane są nieidealnie, to znaczy, że wysokie wyniki na zmiennej niezależnej są przynajmniej częściowo powiązane z innymi

przyczynami niż wysokie wyniki w drugiej zmiennej. W konsekwencji (jeśli zmienne mają rozkład normalny), im wyższy wynik w jednej zmiennej, tym wyższy w nim będzie udział czynników losowych, zawyżających wynik, których powtórzenie się przy pomiarze drugiej zmiennej jest mało prawdopodobne. Tym samym wyniki w tej drugiej zmiennej nie będą tak wysokie.

Zjawisko to można zilustrować (upraszczając nieco związane z tym zagadnienia biologiczne) na przykładzie wzrostu rodziców i dzieci. Wzrost jest cechą silnie związaną z cechami genetycznymi, jednak do pewnego stopnia zależną też od czynników środowiskowych, takich jak odżywianie, tryb życia, aktywność fizyczna itp. Oznacza to, że populacja rodziców o wzroście 190 cm niekoniecznie posiada geny wzrostu o „wartości" 190 cm. W populacji tej są osoby, których „genotypowy wzrost" jest jeszcze większy niż 190 cm, lecz niekorzystne czynniki środowiskowe spowodowały, że ich wzrost fenotypowy wynosi 190 cm. Podobnie w populacji tej znajdą się osoby, których geny opiewają na wzrost niższy niż 190 cm, lecz korzystna konfiguracja czynników środowiskowych podwyższyła wzrost do 190 cm. Pierwsza z tych możliwości zdarzać się będzie rzadziej, ponieważ wzrost jest cechą o rozkładzie normalnym, a osobników wyższych jest mniej niż niższych. Zatem wysoki wzrost jest częściej spowodowany losowymi korzystnymi czynnikami środowiskowymi. Jest mało prawdopodobne, aby taki korzystny układ czynników środowiskowych powtórzył się u dzieci tych osób, które w konsekwencji będą średnio niższe niż rodzice. Zjawisko to ilustruje poniższy wykres:



Wykres 1. Linia regresji wyznaczona dla standaryzowanych wartości dwóch nieidealnie skorelowanych zmiennych

Jak widać, dla osoby, która ma w zmiennej niezależnej wynik równy dwóm odchyleniom standardowym powyżej średniej, wynik przewidywany wynosi tylko niewiele powyżej jednego odchylenia standardowego powyżej średniej. Zjawisko to wystąpi zawsze wtedy, kiedy korelacja między badanymi nie jest idealna (czyli w praktyce zawsze, przynajmniej jeśli chodzi o psychologię). Zawsze bowiem, jeśli korelacja nie jest idealna, linia regresji (wyznaczona na podstawie danych wystandaryzowanych) będzie odchylona od linii regresji idealnej, przebiegającej pod kątem 45 stopni. Regresja do średniej będzie więc tym większa, im słabiej skorelowane są badane zmienne (a tym samym większe odchylenie linii regresji od 45 stopni).

❑ 3. Czy regresja do średniej jest zawsze artefaktem statystycznym?

Odpowiedź na to pytanie jest właściwie przecząca. Regresja do średniej nie zawsze musi być rezultatem artefaktu, jeśli rozumiemy przezeń coś, co wynika wyłącznie z błędów pomiaru i nie ma żadnego odbicia w rzeczywistych czynnikach, istniejących niezależnie od sposobu pomiaru i analizowania danych. Regresja do średniej może być spowodowana interesującymi czynnikami przyczynowymi, w którym to przypadku nie musi być traktowana jako artefakt. Przedstawione dotąd rozważania dotyczyły regresji do średniej spowodowanej błędami pomiaru. Nie jest to jednak jedyny możliwy powód regresji do średniej. Wyróżnić można dwie klasy przyczyn regresji do średniej: (1) związane z błędem pomiaru i (2) związane z czynnikami przyczynowymi oddziaływującymi na badane zmienne. Nesselroade i in. (1980) określają te dwa typy przyczyn regresji do średniej jako stochastyczne oraz deterministyczne. Żeby to zilustrować, rozważmy następujące dwa przykłady.

Karylowski (1985) zaproponował symulację efektu regresji do średniej, której można użyć do zademonstrowania tego efektu i jego przyczyn. Słuchaczom przedstawia się najpierw pojęcie poziomu aspiracji, jako względnie stałej cechy indywidualnej. Wyjaśnia się, że zarówno skrajnie wysokie, jak i skrajnie niskie poziomy aspiracji są nieprzystosowawcze i szkodliwe. Następnie mówi się słuchaczom, iż posiada się pewien szczególny rodzaj zdolności terapeutycznych, które pozwalają zmieniać poziom aspiracji u osób, u których poziom ten jest zbytnio zawyżony lub zaniżony.

Przedstawia się następnie słuchaczom możliwe sposoby pomiaru poziomu aspiracji i wyjaśnia, że nie istnieje coś takiego jak pomiar idealny, pozbawiony błędu. Objaśnia się, że wynik pomiaru będzie zawsze zależał od rzeczywistego poziomu aspiracji badanej osoby oraz od różnych przelotnych przypadkowych zakłócających czynników, zwanych błędem pomiaru. Daje się kilka przykładów przyczyn błędu pomiaru, na przykład

nastrój badanego podczas testowania, mechaniczne pomyłki podczas wypełniania testu, niezrozumienie niektórych pozycji testu itp. Wreszcie prosi się każdego słuchacza, aby pomyślał o trzech czy czterech znanych sobie osobach i dokonał „pomiaru" ich poziomu aspiracji, na sześciostopniowej skali, na której 1 i 2 oznaczają niskie aspiracje (poniżej rzeczywistych możliwości danej osoby), 3 i 4 oznaczają adekwatny do możliwości poziom aspiracji, a 5 i 6 – nierealistyczny, zawyżony poziom aspiracji. Po dokonaniu tego każdej osobie przyporządkowuje się pewien „błąd pomiaru"; dokonuje się tego za pomocą rzutu kostką do gry. „Wynik otrzymany" jest średnią z wartości przyporządkowanej danej osobie przez słuchacza oraz wartości wylosowanej. Na podstawie „wyników otrzymanych" selekcjonuje się skrajne grupy, na przykład górne i dolne 10%. Słuchaczom oznajmiamy, że dzień wcześniej oddziałaliśmy na ich znajome osoby swoją siłą terapeutyczną, po czym następuje procedura retestu. Mówimy słuchaczom, że ponieważ nasze zdolności psychiczne są fikcją, w niniejszym reteście powinni przyporządkować swoim badanym osobom dokładnie takie same wartości aspiracji, jak w preteście. Błąd pomiaru trzeba natomiast powtórnie wylosować, poprzez ponowny rzut kostką. Na koniec porównuje się średni poziom aspiracji w obu grupach w pierwszym „testowaniu" i w drugim. Słuchacze zauważą natychmiast obniżenie się średniej w grupie o wysokiej aspiracji i podwyższenie średniej w grupie o niskiej aspiracji (Karylowski, 1985). Zazwyczaj wystarczy kilkanaście obserwacji, aby uzyskać zmianę wyników statystycznie istotną.

Demonstracja ta jest przykładem sytuacji, w której za całość efektu regresji do średniej odpowiada wyłącznie błąd pomiaru. Pokazuje ona dobitnie, jak pod całkowitą nieobecność zmiany wyniku prawdziwego pojawi się zmiana wyniku otrzymanego. Rozważmy drugi przykład, w którym, przeciwnie do poprzedniego, za regresję do średniej odpowiadają nie błędy pomiaru, lecz przyczyny niezależne od artefaktów obliczeniowych. Pewien badacz zaprojektował nową metodę zmiany osobowości dzieci agresywnych, a konkretnie – metodę zmniejszania agresywności u dzieci agresywnych. Badacz ten jest zorientowany behawioralnie w tym sensie, że agresywność, jako cecha osobowości, jest dla niego równoznaczna z widocznymi przejawami agresji – agresywne jest to dziecko, które zachowuje się agresywnie. Pomijając chwilowo zagadnienie, czy taka konceptualizacja agresywności jest zasadna, stwierdzić należy, że w jej wypadku trudno mówić o błędach pomiaru, ponieważ liczba zachowań agresywnych nie jest tu traktowana jako pośredni obserwowalny wskaźnik pewnej nieobserwowalnej „latentnej" cechy osobowości, zwanej agresywnością. Przeciwnie, liczba zachowań agresywnych jest agresywnością. Skoro tak, badacz ten nie ma problemów z błędem pomiaru, ponieważ liczbę zachowań agresywnych w jednostce czasu można określić praktycznie bezbłędnie. Pomimo to, dzieci agresywne okażą się najpraw-

dopodobniej po zakończeniu postępowania korekcyjnego mniej agresywne, nawet jeśli postępowanie to było całkowicie nieskuteczne.

Przyczyną regresji do średniej w tym przypadku nie są błędy pomiaru, lecz czynniki przyczynowe oddziałujące na badane zmienne w obu pomiarach. Wysoka liczba aktów agresji w pewnej grupie dzieci mogła być spowodowana określoną konfiguracją wyznaczników agresji, jaka zdarzyła się w momencie pomiaru. Ogólnie mówiąc, agresywność dzieci może być funkcją takich czynników, jak zdarzenia przykre lub przyjemne, jakie spotkały badane dzieci przed momentem pomiaru, sytuacja w domu rodzinnym, jaka panowała przed pomiarem, ewentualne kary za niewłaściwe zachowanie, nastrój dziecka tego danego dnia, spowodowany fluktuowaniem czynników biopsychicznych, i wiele innych. Są to elementy przyczynowe, ale ich konfiguracja w danym momencie u poszczególnych osobników jest bardzo różna – u części korzystna, zmniejszając liczbę aktów agresji, u części obojętna, u części wreszcie niekorzystna, powiększając liczbę aktów agresji. W momencie pomiaru badacz trafia na pewną „zastaną" konfigurację wyznaczników agresji u wszystkich dzieci. Jeśli wybierze dzieci przejawiające dużo aktów agresji, to należy oczekiwać, że w drugim pomiarze liczba aktów agresji u tych dzieci będzie średnio mniejsza, ponieważ jest mało prawdopodobne, aby u wszystkich powtórzyła się ta szczególna konfiguracja wyznaczników agresji, jaka zdarzyła się za pierwszym razem. W konsekwencji, zaobserwowana zostanie zmiana wyników (najprawdopodobniej istotna statystycznie) zgodna z postawioną hipotezą, lecz niedowodząca jednak skuteczności zastosowanego postępowania korekcyjnego.

W omawianym przypadku regresja nie jest spowodowana tylko błędami pomiaru, lecz oddziaływaniem czynników ważnych dla agresywności, potencjalnie identyfikowalnych. Ewentualna pomyślna identyfikacja tych czynników może stanowić wartościowy wkład w wiedzę o agresywności. Czy zatem regresja do średniej stanowi w tym przypadku jakikolwiek problem? Tak, problem pozostaje, ponieważ zmiana wyników w drugim pomiarze może zostać błędnie przypisana manipulacji eksperymentalnej (Hopkins, 2000). Badacz obserwujący zmniejszenie się liczby aktów agresji po swojej terapii może uznać, że zmniejszenie to jest spowodowane postępowaniem korekcyjnym, tymczasem w rzeczywistości może ono wynikać z czynników, być może interesujących, lecz całkowicie niezależnych od tej terapii. Innymi słowy, gdyby korekcji w ogóle nie zastosowano, do zmniejszenia się liczby aktów agresji doszłoby również.

Można też zadać pytanie, jaki jest wpływ zjawiska regresji do średniej na jakość predykcji, do której wykorzystywana jest analiza regresji. Wartość zmiennej zależnej zawsze jest bliższa średniej niż wartość predyktora. Czy to znaczy, że wszelkie predykcje obciążone są błędem, polegającym na tym, że wartości oczekiwane są zawsze zaniżone lub zawyżone w stronę średniej? Innymi słowy, czy analiza regresji, mająca na celu

predykcję, przynosi zawsze wyniki obciążone? Odpowiedź jest przecząca. Istnieje mnóstwo zagrożeń dla precyzji predykcji, poczynając od niedokładności pomiaru badanych zmiennych, lecz pozostaje faktem, że predykowana dla zmiennej zależnej wartość, bliższa średniej niż wartość predyktora, jest dobrą predykcją! Obciążona, niedokładna i błędna byłaby właśnie predykcja zakładająca, że wartość zmiennej zależnej będzie odległa o tyle samo odchyleń standardowych od średniej co wartość predyktora. Innymi słowy, matematyka regresji do średniej jest tylko odbiciem pewnego zjawiska realnego, polegającego na tym, że dla dwóch niedoskonale skorelowanych zmiennych, przewidywana z pewnej wartości jednej z tych zmiennych wartość drugiej zmiennej jest bliższa średniej niż wartość, dla której dokonano predykcji. Regresja do średniej bywa traktowana jako „trywialna matematyczna tautologia" (Gottman i Rushe, 1993), pozostaje jednak faktem, że jest ona zjawiskiem wpisanym w rzeczywistość istniejącą dookoła nas, w dodatku na tyle silnym, że często zauważalnym nawet bez analiz statystycznych. Dzieci geniuszy, choć zdolne, rzadko są geniuszami; sportowcy najlepsi w danym sezonie zwykle nie są najlepsi w następnym, sequel przeboju kinowego jest zazwyczaj gorszy od pierwowzoru, studenci dobrzy w pierwszym semestrze nie są tak dobrzy w następnym, dzieci wybitnych aktorów rzadko zachodzą tak daleko, jak ich rodzice; partie polityczne hołubione w jednych wyborach przepadają w następnych; dziesiątki tego rodzaju przykładów można znaleźć w pracach dotyczących regresji do średniej (np. Campbell i Kenny, 1999) lub zaobserwować samemu. „Po rzeczach niezwykłych najczęściej następują zwyklejsze", jak ujęli to Shaughnessy i Zechmeister (1990).

❑ 4. Inflacja wyobraźni – fakt czy artefakt?

Rozważania o regresji do średniej warto zilustrować pewnym przykładem, a mianowicie dyskusją nad zjawiskiem „inflacji wyobraźni". Przykład ten jest wartościowy, między innymi dlatego, że ilustruje trudności, na jakie trafiają badacze, usiłując ustalić, czy w danym przypadku za uzyskane wyniki odpowiada wyłącznie regresja do średniej, czy też zmienne badane w danym eksperymencie.

Termin „inflacja wyobraźni" (*imagination inflation*) został użyty po raz pierwszy w 1996 roku przez Garry, Manninga, Loftus i Shermana, na określenie zjawiska polegającego na zwiększeniu subiektywnej pewności autentyczności pewnego nieprawdziwego wspomnienia, wskutek wyobrażania sobie zdarzenia, które jest treścią tego wspomnienia. W swoim eksperymencie Garry i in. (1996) najpierw poprosili badanych o podanie, jak pewni są oni, że w ich dzieciństwie miały miejsce różne zdarzenia, wśród których były też mało prawdopodobne, jak na przykład utknięcie na drzewie albo wybicie okna ręką, na skali od 1 – „jestem całkowicie pewny,

że nie zdarzyło się", do 8 – „jestem całkowicie pewny". Na podstawie tego pierwszego pomiaru do dalszych badań wybrano osiem zdarzeń, których średnie oceny wskazywały na to, że prawdopodobnie nie miały one miejsca. Dwa tygodnie później osoby badane zostały poproszone o intensywne wyobrażanie sobie tych zdarzeń, po czym powtórnie oszacowały swoją subiektywną pewność prawdziwości ich wystąpienia w dzieciństwie. Garry i in. (1996) stwierdzili, że u badanych częściej dochodziło do zwiększenia niż do zmniejszenia pewności wystąpienia zdarzenia po drugim badaniu. Zjawisko to nazwali „inflacją wyobraźni" i zinterpretowali jako eksperymentalną demonstrację, jak procedury wyobrażeniowe mogą doprowadzać do kreowania wspomnień. Zjawisko „inflacji wyobraźni" wzbudziło duże zainteresowanie i bywało z powodzeniem replikowane, w każdym razie na próbach ludzi młodych (np. Heaps i Nash, 1999; Paddock, Joseph, Ming Chan, Terranova, Manning i Loftus, 1998).

Jednak w 2001 roku Pezdek i Eddy przedstawiły artykuł, w którym dowodziły, że zjawisko inflacji do średniej może być z powodzeniem wytłumaczone za pomocą regresji do średniej, a nie jako przykład działania wyobraźni na pamięć. Główne argumenty Pezdek i Eddy (2001) były następujące:

1. Do zwiększenia pewności wystąpienia zdarzenia doszło *zarówno* w wypadku zdarzeń wyobrażanych, jak i zdarzeń kontrolnych, których badani sobie nie wyobrażali.

2. Pezdek i Eddy (2001) uzyskały od Garry i in. (1996) ich oryginalne dane, nieprzedstawione w artykule, dotyczące zdarzeń, których badani w pierwszym teście byli pewni. W wypadku zdarzeń, których wystąpienia badani byli w pierwszym teście pewni, doszło do *zmniejszenia* tej pewności w drugim badaniu. Wyniki są więc zgodne ze wzorcem regresji do średniej – niskie wyniki uległy powiększeniu, a wysokie zmniejszeniu.

3. Bezpośredni test istotności statystycznej różnicy między zdarzeniami wyobrażanymi i niewyobrażanymi, oparty na odsetkach osób badanych, nie był możliwy, lecz Garry i in. (1996) wykonali test istotności, wykorzystujący jako jednostki obserwacji poszczególne zdarzenia, a nie osoby badane. Uzyskali różnicę istotną statystycznie – częściej zmianom ulegały zdarzenia wyobrażane. Jednak Pezdek i Eddy (2001) argumentują, że analiza tego rodzaju oparta jest tylko na liczbie zmian i ignoruje *wielkość* zmiany. Na przykład, jeśli trzy osoby zmieniły swoją odpowiedź z „2" na „3", a dwie zmieniły z „4" na „1", to w świetle analiz Garry i in. (1996) dochodziło częściej do zwiększenia pewności, lecz w rzeczywistości średnia pewność uległa zmniejszeniu, a nie zwiększeniu. Analiza wykonana przez Pezdek i Eddy (2001), oparta na wielkościach efektu, a nie tylko na liczbie zmian, wykazała brak istotności statystycznej różnicy pomiędzy zdarzeniami wyobrażanymi i niewyobrażanymi.

4. Pezdek i Eddy (2001) wykonały własne badania, naśladujące dosyć wiernie eksperyment Garry i in. (1996). Uzyskały wyniki identyczne: do

zwiększenia pewności doszło zarówno w przypadku zdarzeń wyobrażanych, jak i niewyobrażanych, różnica była istotna statystycznie, jeśli brano pod uwagę liczbę zmian, lecz nie była istotna, kiedy analizowano wielkość zmiany. Co więcej, Pezdek i in. (2001) stwierdziły, że wielkość zmiany zależy od skrajności pierwszego pomiaru – im był on bardziej skrajny, tym efekt zmiany większy (dotyczyło to zarówno itemów o wysokiej, jak i o niskiej pewności, oraz zarówno itemów wyobrażanych, jak i niewyobrażanych).

Pezdek i Kelly (2001) konkludują, że wyniki te są zgodne ze wzorcem predykowanym przez regresję do średniej, a hipoteza inflacji wyobraźni jest do ich wyjaśnienia niepotrzebna.

W odpowiedzi na te zarzuty Garry, Sharman, Wade, Hunt i Smith (2001) nie zgodzili się, że za „inflację wyobraźni" odpowiedzialna jest regresja do średniej. Garry i in. (2001) zgodzili się, że regresją do średniej można tłumaczyć podwyższenie się pewności w wypadku zdarzeń niewyobrażanych, lecz nie w wypadku zdarzeń wyobrażanych. Broniąc swoich wyników, podważali oni zasadność analiz regresji i wariancji, zastosowanych przez Pezdek i Eddy (2001), na przykład stwierdzili, że analiza wariancji jest wątpliwa w sytuacji, kiedy badane zmienne mają silnie skośny rozkład, a taki właśnie mają w eksperymentach nad inflacją wyobraźni. Bronili też swojej metody wyznaczania istotności statystycznej różnicy między zdarzeniami wyobrażanymi i niewyobrażanymi, opartej na zliczaniu zmian, a nie analizowaniu wielkości zmian.

Brak tu miejsca na przedstawienie wszystkich szczegółów tego pasjonującego sporu i wszystkich argumentów obu stron, nie chcę też powiedzieć, że, moim zdaniem, Pezdek i Eddy (2001) udowodniły, że „inflacja wyobraźni" jest skutkiem regresji do średniej. Bezsporne jednak wydaje się, że interpretacja eksperymentów dotyczących „inflacji wyobraźni" w duchu regresji do średniej jest możliwa. Jest to dobry przykład pewnej sytuacji patowej – wyniki eksperymentu *mogły* być spowodowane manipulacjami eksperymentalnymi, ale równie dobrze *mogły* też być w całości skutkiem regresji do średniej. Z uwagi na trudności z pomiarem kontrolnym, wątpliwości takie nie mogły zostać jednoznacznie rozstrzygnięte, a każda ze stron pozostała przy swojej opinii.

❑ 5. Jak ustrzec się regresji do średniej

Regresja do średniej jest zjawiskiem dotyczącym obserwacji o skrajnych wynikach. Najprostszy sposób uniknięcia jej polegałby zatem na rezygnacji z analizy grup skrajnych. Jeśli badacz zamierza na przykład analizować zmianę cech osobowości wskutek pewnej manipulacji, to może zmierzyć u *wszystkich* badanych tę cechę, następnie poddać połowę badanych manipulacji, która jest przedmiotem badania, po czym powtórnie zmie-

rzyć analizowaną cechę u *wszystkich* badanych. Analiza wariancji dla interakcji powtórzonego pomiaru z czynnikiem międzygrupowym odpowiedziałaby na pytanie, czy zmiana w zakresie zmiennej zależnej jest inna dla grupy eksperymentalnej niż dla kontrolnej.

Procedura taka jest jednak często wykluczona, ponieważ przedmiotem badania są z założenia osoby odznaczające się skrajnym natężeniem badanej zmiennej. Na przykład, w opisywanych powyżej badaniach nad inflacją wyobraźni, Garry i in. (1996) nie mogli uwzględnić wszystkich badanych, ponieważ „inflacja wyobraźni" miała z definicji dotyczyć zdarzeń niebyłych, a jedyną gwarancją niewystąpienia zdarzenia była pewność jego niewystąpienia w dzieciństwie, przejawiająca się niskimi wynikami na skali pewności w pierwszym pomiarze.

W praktyce wydzielenie grupy skrajnej jest niezbędne zawsze wtedy, kiedy przedmiotem badania mogą być tylko obserwacje odbiegające od średniej; jest to na przykład niemal regułą w badaniach nad skutecznością (psycho)terapii, adresowanej przecież do osób dotkniętych zaburzeniami, zatem z definicji stanowiących pewną nietypową i mniej od osób zdrowych liczną podpopulację. W takich przypadkach wskazane jest zastosowanie metod redukujących czy kontrolujących zjawisko regresji do średniej. Wyróżnić można dwie sytuacje: kiedy możliwe jest zastosowanie grupy kontrolnej oraz – kiedy nie jest to możliwe.

Grupa kontrolna. Nie będzie zapewne zbyt odkrywczym stwierdzenie, że najlepszym sposobem poradzenia sobie ze zjawiskiem regresji do średniej jest rozlosowanie osób badanych do dwóch grup: eksperymentalnej oraz kontrolnej, w której nie stosuje się manipulacji stanowiącej przedmiot badań, słowem – zastosowanie procedury eksperymentalnej w pełnym tego słowa znaczeniu. Plan eksperymentalny będzie zatem zawierał jeden dwupoziomowy czynnik międzygrupowy (grupa eksperymentalna oraz kontrolna) oraz jeden dwupoziomowy czynnik z powtarzanymi pomiarami (pomiar pierwszy oraz drugi). W obu grupach nastąpi zbliżenie się wyników drugiego pomiaru do średniej. Jeśli celem manipulacji było podwyższenie wyników, to jej ewentualna skuteczność przejawi się większym wzrostem wyników w grupie eksperymentalnej niż kontrolnej. Najprostszym sposobem sprawdzenia tego jest analiza wariancji dla interakcji między czynnikiem grupowym a powtórzonym pomiarem.

Mówiąc o grupie kontrolnej, warto zauważyć, że niecelowe jest wydzielanie grup skrajnych (z dodatkiem ewentualnie grupy „środkowej") i traktowanie ich jako nawzajem dla siebie kontrolnych. Jeśli badacz zamierza sprawdzić hipotezę, że jego manipulacja podwyższy wyniki osobom o niskich wynikach, a obniży je osobom o wysokich wynikach, to poprawny jest taki plan eksperymentalny, w którym występują dwie grupy badanych: eksperymentalna, poddana manipulacji, i kontrolna. Analiza polega na porównaniu testu i retestu w obu tych grupach, po wydzieleniu w obydwu z nich podgrup skrajnych. W obu grupach – kon-

trolnej oraz eksperymentalnej – pojawi się obniżenie wyników wysokich i podwyższenie niskich, a potwierdzenie się swej hipotezy badacz uzyska wtedy, gdy to obniżenie i podwyższenie się wyników będzie istotnie statystycznie większe w grupie eksperymentalnej niż kontrolnej.

Kontrola statystyczna. Wszędzie tam, gdzie jest to możliwe, stosowanie odpowiedniej grupy kontrolnej lub porównawczej jest wymogiem zasadniczym. Na przykład poważni analitycy nie zwracają już obecnie uwagi na badania nad skutecznością psychoterapii pozbawione takiej grupy (Chambless i Ollendick, 2001). Problem pojawia się jednak, kiedy badania nie mają i nie mogą mieć charakteru kontrolowanego eksperymentu, lecz są oparte na danych obserwacyjnych. Jest to typowe na przykład dla analiz ekonomicznych, które w konsekwencji są szczególnie narażone na błędy wynikające z niezauważenia czy niedocenienia efektu regresji do średniej (Friedman, 1992), lecz także dla wielu obszarów psychologii, na przykład longitudinalnych obserwacji zmiany pewnej cechy osobowości w czasie. Problem ten napotka na przykład badacz rozwoju człowieka, interesujący się tym, czy dzieci zaburzone mają tendencję do powracania do normy w miarę upływu lat. Badacz ten odkryje zapewne, że odpowiedź jest twierdząca i będzie musiał zmierzyć się z interpretacją opartą na regresji do średniej.

Kiedy zastosowanie grupy kontrolnej nie jest możliwe, godne uwagi jest zalecenie, aby badani byli wybierani do grup skrajnych na podstawie kilku testów, nie tylko jednego, jak również, żeby pomiar końcowy stanowił wypadkową kilku testów, nie tylko jednego (Yudkin i Stratton, 1996). Z podobnych powodów, dla których istnieje regresja do średniej (losowej fluktuacji błędów), mniej prawdopodobne jest, aby dana osoba trafiła do grupy skrajnej z powodu błędów pomiaru, kiedy jest selekcjonowana na podstawie kilku testów, niż wtedy, kiedy jest selekcjonowana na podstawie jednego testu. Zastosowanie kilku testów nie da wprawdzie gwarancji zlikwidowania regresji do średniej, lecz zmniejszy prawdopodobieństwo jej wystąpienia. Metoda ta jednak będzie nieskuteczna w sytuacji, kiedy regresja do średniej nie została spowodowana błędami pomiaru.

Poza oparciem pomiaru na kilku testach, w pewnych sytuacjach możliwe do zastosowania są statystyczne metody kontroli regresji do średniej. Są one zbliżone w swej idei do metod opartych na grupie kontrolnej. Ich istotą jest szacowanie, jak silny efekt regresji do średniej jest oczekiwany w danym wypadku, a następnie porównywanie tego oczekiwanego efektu z rzeczywistą zmianą wyników. Jeśli z obliczeń wynika, że zbliżenie do średniej jest większe niż oszacowane zbliżenie wynikające z regresji do średniej, to zyskujemy mocniejszy grunt dla argumentacji, że zaobserwowana w badaniach zmiana nie jest tylko artefaktem statystycznym. Stosowanie tych metod ogranicza się jednak do sytuacji, kiedy znane są właściwości psychometryczne metod używanych w badaniach. Znana jest

na przykład korelacja retestowa testu, wyznaczona na podstawie niezależnych badań (Hopkins, 2000), która może posłużyć do oszacowania zmiany wyników, oczekiwanej na podstawie regresji do średniej, a następnie porównania jej ze zmianą uzyskaną w danych badaniach. Jeśli nic nie wiemy o właściwościach psychometrycznych metod użytych w badaniach, to metody kontroli statystycznej nie mają zastosowania lub mają je tylko w bardzo ograniczonym zakresie, ale pozwalają one na przykład na identyfikację tych badanych, w wypadku których zmiana wyników jest na tyle duża, że trudno wyjaśnialna w kategoriach regresji do średniej (Hsu, 1995).

Dobrym pomysłem jest też badanie zmiany wstecz – jeśli na przykład badacz wykrył, że osoby gorsze w preteście są lepsze w postteście, to może on sprawdzić, czy osoby gorsze w postteście są lepsze w preteście. Jeżeli tak się właśnie stanie, to badacz ten powinien podejrzewać, że za jego wyniki odpowiedzialna jest regresja do średniej (Campbell i Kenny, 1999).

Campbell i Kenny (1999) proponują też różne metody graficznej wizualizacji danych, które mogą pomóc zorientować się, czy za uzyskane rezultaty odpowiedzialna jest regresja do średniej. Ogólnie jednak mówiąc, statystyczne metody kontroli przynoszą bardziej dyskusyjne wyniki niż stosowanie grupy kontrolnej, nie są wolne od poważnych słabości, a nieumiejętne ich stosowanie może zakończyć się tworzeniem nowych artefaktów, zamiast zwalczeniem artefaktów powodowanych przez regresję do średniej (Campbell i Kenny, 1999).

W 1933 roku amerykański statystyk i ekonomista Horace Secrist, postawiony przed zadaniem wyjaśnienia przyczyn Wielkiego Kryzysu, napisał książkę *The triumph of mediocrity in business*, w której za pomocą ponad 200 tabel i wykresów „dowiódł", że dobre przedsiębiorstwa w miarę upływu czasu mają gorsze osiągnięcia, a złe stają się lepsze, i przepowiedział, że wskutek tego procesu wkrótce wszystkie przedsiębiorstwa będą miały takie same, średnie osiągnięcia. Obecnie, po 70 latach, wiemy, że nigdzie na świecie nie obserwuje się uśredniania się rozmiarów czy dochodów przedsiębiorstw. Miażdżącą recenzję książki Secrista napisał znany statystyk Hotelling (1933; za: Stigler, 1997) i od tego czasu przypadek ten pozostaje jednym ze sztandarowych przykładów błędów spowodowanych nieuwzględnieniem regresji do średniej. Warto jednakże zdać sobie sprawę, kim była osoba, której zarzucono tak szkolny błąd: Horace Secrist był profesorem ekonomii, statystykiem pracującym dla amerykańskiej World Industry Commision, dyrektorem Bureau of Business Research w Northwestern University, Kierownikiem Badań w Claremont College, członkiem Commision of Industry Relations, statystykiem-superwizorem w U.S. Railroad Labor Board, członkiem American Statistical Association, członkiem Association of University Professors,

Manchester Statistics Society, członkiem stowarzyszenia Phi Beta Kappa i jednym z najbardziej znanych statystyków swoich czasów (dane z internetowej wersji leksykonu *Who's Who in America*, 1935). Jak z tego wynika, nie trzeba koniecznie być kiepskim badaczem, żeby wpaść w pułapkę regresji do średniej.

Campbell i Kenny (1999) stwierdzili, że uniwersalności występowania zjawiska regresji do średniej towarzyszy równie uniwersalne jego niezrozumienie, przejawiane zarówno przez niefachowców, jak i przez niektórych ekspertów w dziedzinie statystyki, w tym kilku noblistów. Streiner (2001) dodał, że regresja do średniej jest zjawiskiem powszechnym, lecz, jak to ma miejsce w wypadku wielu zjawisk powszechnych, mamy tendencję do unikania jej brzydoty, czujemy się zakłopotani, mówiąc o niej publicznie i ciągle dziwimy się jej ogromnej sile. Stigler (1997) napisał, że odkrycie przez Galtona regresji do średniej pozostaje jednym z największych triumfów w całej historii statystyki oraz historii nauki, triumfem, który każda generacja musi od nowa uczyć się doceniać, i który dotyczy zjawiska najwyraźniej nigdy nietracącego swej zdolności sprawiania niespodzianek.

Jeśli stwierdzenia te wydają się przesadzone, to końcowe uwagi niniejszego artykułu można sprowadzić do dwóch zaleceń, zaczerpniętych z literatury przedmiotu: po pierwsze, jeśli zobaczymy jakiś interesujący wynik empiryczny dotyczący zmiany, a oparty na badaniach na wyselekcjonowanych skrajnych grupach, to regresja do średniej powinna być wyjaśnieniem standardowym, które musi zostać najpierw odrzucone, zanim zaproponuje się i zacznie poważnie traktować wyjaśnienia merytoryczne (Senn, 1997). Jeśli interpretacja danego efektu w kategoriach regresji do średniej nie może zostać odrzucona, to chociaż nie przesądza to jednoznacznie o fałszywości interpretacji merytorycznej, zdrowy rozsądek nakazuje traktować tę ostatnią ostrożnie. Po drugie, trzeba zdawać sobie sprawę, że jeśli kiedykolwiek rozpoczniemy badania nad zmianą, nie mając możliwości zastosowania grupy kontrolnej, to uczynimy tak na własne ryzyko – ryzyko nierozwiązywalnej niekonkluzywności uzyskanych wyników.